Vol.2019-ARC-237 No.9
2019/7/24

パケット処理キャッシュにおけるパイプライン化とマルチポート化の評価

田中 京介[1,a)]　八巻 隼人[1]　三輪 忍[1]　本多 弘樹[1]

概要：インターネットトラフィックは年々急激に増加しており，近い将来には 1Tbps の超大容量回線が登場することが見込まれている．しかしながら，既存のインターネットルータのパケット処理スループットは 400Gbps 程度が限界であり，1Tbps 達成のためには大幅なスループットの向上が必要とされる．これに対し，インターネットルータのパケット処理においてスループット上のボトルネックとなっているテーブル検索を，キャッシュを用いて高速化するパケット処理キャッシュ (PPC: Packet Processing Cache) が提案されている．PPC を用いたテーブル検索処理では，PPC のキャッシュミス率がスループットを決定することから，これまでキャッシュミスを削減する様々な研究が行われてきた．一方で，マイクロプロセッサ等のキャッシュに用いられるパイプライン化，マルチポート化といった手法を PPC に適用した場合の，スループットや消費電力に与える影響は議論されていなかった．そこで本報告では，これまで報告してきた階層化に加え，パイプライン化とマルチポート化，あるいはそれらの手法の組み合わせが PPC に及ぼす影響について評価した．評価の結果，3 手法を組み合わせることにより，階層化のみを適用した場合と比較して消費電力を 63.9%改善した上で，1Tbps 超のスループットを達成できることが分かった．一方で，これらの手法は本質的に初期参照ミスが大きくなるネットワークトラフィックでは効果が低く，さらなる性能改善のためには初期参照ミスの削減に取り組む必要があることを示した．

キーワード：ネットワークルータ，パケット処理，キャッシュ，メモリ階層

1. はじめに

インターネットトラフィックは年々増加しており [4]，近い将来には 1Tbps 級の超大容量回線が必要となることは明らかである [6]．しかしながら，既存のインターネットルータのパケット処理スループットは 400Gbps 程度が限界であり，1Tbps 達成のためにはスループットを 2.5 倍向上する必要がある．それに加え，近年のインターネットルータ設計においては，製造および運用コストの観点から消費電力とチップ面積の削減が強く求められている．特に消費電力は転送するパケット数に比例して増加することから，1Tbps 級のパケット処理性能を現実的な消費電力で実現するインターネットルータが必要とされている．これらの設計制約を満たすためには，ルータのアーキテクチャを大幅に変更する必要がある．

インターネットルータにおいては，パケットによるテーブル検索処理がスループットを向上する上でのボトルネックとなっている [5]．テーブル検索処理では，ルータへ到着したパケットの転送に要する情報（宛先，転送可否，優先度等）を，ルータ内のテーブルを検索することによって得る．一般的なインターネットルータのテーブル検索処理では，ルーティングテーブル，ARP (Address Resolution Protocol) テーブル，ACL (Access Control List)，そして QoS (Quality of Service) テーブルといった複数テーブルの検索を要する．

近年のハイエンドルータは，これらのテーブル 1 サイクルで検索可能な高速メモリである TCAM (Ternary Content Addressable Memory) へ格納することにより，高いテーブル検索性能を実現している．しかしながら，現在の TCAM の検索性能では，最小サイズのパケットが連続してルータへ到着するワーストケースにおいて 100Gbps 程度のスループットが限界であり，TCAM であっても将来的な回線速度に対する検索性能の不足が懸念される．加えて，TCAM は一般的に使用されるメモリである SRAM (Static Random Access Memory) に比べ，1 回のアクセスに必要な消費電力が極めて大きく，ルータの全消費電力の 40%程度を TCAM が占めているとの報告もある [8], [11].

[1] 電気通信大学
1-5-1, Chofugaoka, Chofu, Tokyo 182–8585, Japan
[a)] kyontan@hpc.is.uec.ac.jp

テーブル検索処理を高速化する手法として，小容量かつ高速な SRAM を TCAM のキャッシュとして用いるパケット処理キャッシュ (Packet Processing Cache: PPC) が提案されている．インターネット上では，1 つの送信元から 1 つの宛先へ短時間に大量のパケットを送信することが多く，パケットを区別するフローには時間的局所性がある．そのため，TCAM の検索結果をキャッシュに記憶して再利用することにより，TCAM のアクセス回数を削減し，高速かつ低消費電力なテーブル検索処理を実現できる．

PPC の利用によってインターネットルータのスループットと消費電力は大幅に改善するが，これまでに発表された PPC 搭載インターネットルータのスループットは数百 Gbps 程度に留まる．これは，後述するように，PPC のアーキテクチャの最適化が不十分なためである [15].

本報告では，マイクロプロセッサのキャッシュのスループットを改善する複数の手法を PPC へ適用し，それらの効果を評価する．スループットの改善をするにあたり，我々は PPC の**階層化**，**パイプライン化**，そして**マルチポート化**の 3 手法を選択する．これらの手法は直交するため，これらの手法を組み合わせた場合についても同様に検証する．検証においては，いくつかの設計制約を仮定し，PPC を有するインターネットルータについて，スループット，消費電力および面積の観点から評価を行う．

我々は，これまでに階層化が PPC に与える影響について報告してきたが [13]，本報告はパイプライン化やマルチポート化，あるいはそれらを組み合わせた場合について包括的な評価を行うものである．

本報告における主な貢献は以下の通りである．

- マルチポート化，パイプライン化単体では 1Tbps の達成は難しいことを明らかにした．
- 1Tbps を達成する構成においては，3 手法を組み合わせることにより，階層化単体に比べ消費電力を 64.4%削減できることを明らかにした．
- 70%の面積増加を許容する制約下においては，従来の PPC に対して消費電力を 8.0%削減しつつ 3.55 倍のスループット (1.05Tbps) を達成できることを示した．

本報告は次のような構成になっている．まず第 2 章において PPC をより詳細に説明する．第 3 章において，従来の PPC のボトルネックを分析し，さらなる最適化の必要性を示す．マイクロプロセッサのキャッシュで用いられているスループット向上手法は第 4 章で述べる．実験手法と結果はそれぞれ第 5 章および第 6 章で説明する．関連研究は第 7 章で説明し，結論を第 8 章で述べる．

2. パケット処理キャッシュ (PPC)

図 1 に PPC を用いたテーブル検索処理の概要を示す．既存のルータでは，パケットをどこへ/どのように転送するのかを決定するため，ルーティングテーブルなどの複数のテーブルをオフチップの TCAM を用いて検索する．テーブル検索処理では，パケットヘッダ内の 5 タプル（送信元 IP アドレス, 送信先 IP アドレス, 送信元ポート, 送信先ポート, プロトコル番号）の一部分あるいは全てをキーとして用いて各テーブルの検索を行う．テーブル検索処理によって得られた結果を元に，ルータはあるインタフェースから他のルータへパケットを転送する．

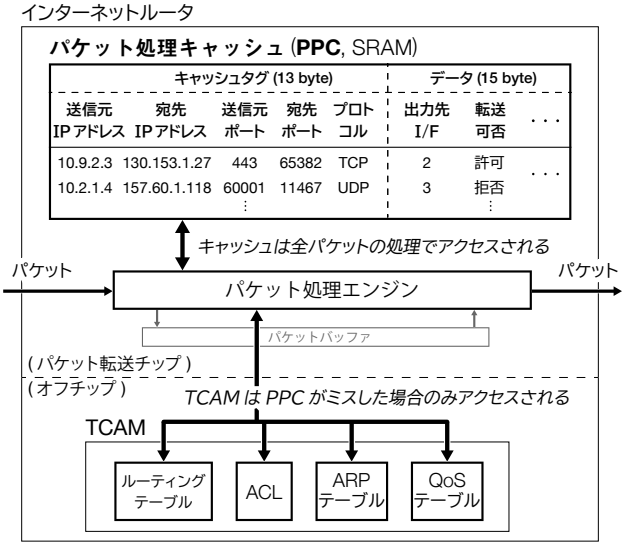


図 1: PPC を用いたテーブル検索処理

PPC は複数の TCAM アクセスを 1 回のキャッシュアクセスにより代替することで，TCAM アクセス数を削減している．具体的には，PPC は 1 パケットの処理に要する複数の TCAM アクセス結果を，パケットの 5 タプルをタグとしてキャッシュの 1 エントリへ格納することによりこれを実現する．一般的な構成のインターネットルータでは，1 つのパケットに対して，出力先ポートを表す 1 バイトの情報，送信元・宛先 MAC アドレスを表す 12 バイトの情報，ACL によるフィルタリング結果を表す 1 バイトの情報，QoS を表す 1 バイトの情報の計 15 バイトのデータを TCAM から取得する．そのため，13 バイトの 5 タプルとあわせ，PPC の 1 エントリあたりの容量は 28 バイトである．

PPC 上にエントリが作成されると，TCAM の代わりに PPC を検索することによって，ルータはテーブル検索処理の結果を取得できる．なお，PPC のインデクスには 5 タプルの CRC ハッシュを用いることが多い．PPC は小容量の SRAM で構成されるため，TCAM に比べ高いスループットを省電力で実現することができる．

従来の PPC は，1K エントリ（28KB），1 階層，1 リードポート，パイプライン化無しの，小容量かつ簡単な構成のキャッシュを用いていた．しかし，そのような小容量の PPC では TCAM へのアクセス率を大きく削減することは困難であった．例えば，最新の研究成果によると PPC のミス率は 14.6%にのぼる [15]. 1Tbps のパケット処理性能

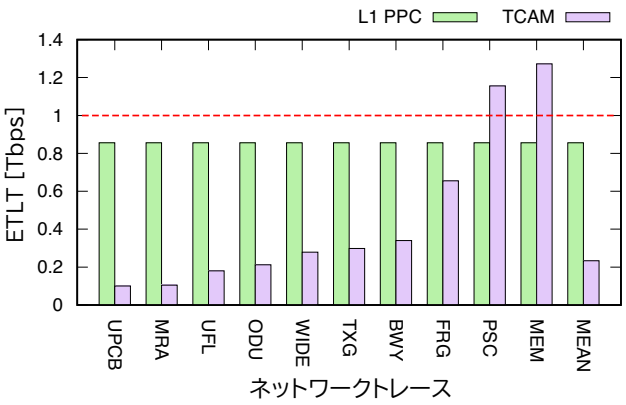


図 2: 従来の PPC と TCAM の ETLT

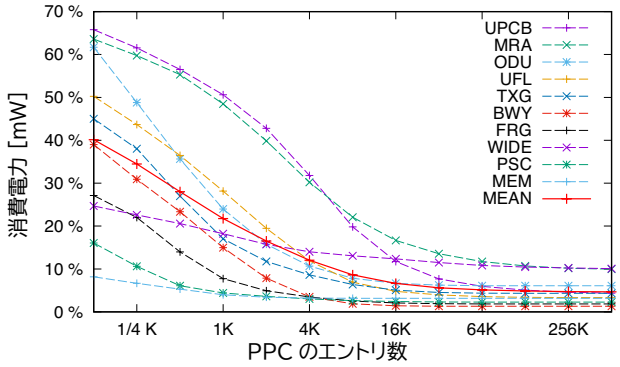


図 3: PPC のエントリ数によるミス率の変化

を実現するためには，次節で詳しく述べるように，スループットを維持しつつ PPC の容量を増やす必要がある．

3. 大容量化と高スループット化の必要性

我々は従来の PPC（1 階層，1K エントリ，1 リードポート，パイプライン化無し）におけるスループットを分析し，1Tbps を達成する上でのボトルネックを明らかにしている [13]．以下ではその概要を述べる．

図 2 は，後述する PPC シミュレータおよび CACTI 6.5 を用いて 10 種の実ネットワークトレースに従来の PPC を適用した場合の，PPC と TCAM の ETLT を示している．ETLT (Effective Table-Lookup Throughput) とは，その階層のメモリがパケットをロスすることなく処理可能なルータの入力トラフィック容量である．また，ルータのスループットは，最も ETLT の低いメモリによって律速される．

図 2 より，まず，従来の PPC は ETLT が低く，いかにキャッシュミスを削減しても 1Tbps を達成できないことがわかる．これはすなわち，PPC をパイプライン化/マルチポート化する，あるいは，PPC の容量を減らすことによって，ETLT を改善する必要があることを意味する．更に，多くのトレースにおいて TCAM の ETLT が不十分であることがわかる．これは，従来の PPC では十分なキャッシュヒットが得られておらず，TCAM アクセスがスループット上のボトルネックとなっていることを示している．したがって，容量の増加等の手段により PPC ミスを削減し，TCAM の ETLT を 1Tbps まで向上させる必要がある．

図 3 は，1 階層の PPC の容量を増加させた時のミス率を示したグラフである．図に示されているように，容量を増加させた際の PPC ミス率の減少傾向はネットワークにより大きく異なるものの，従来の PPC（1K エントリ）よりもエントリ数を増やすことによって PPC ミスは大きく削減できることがわかる．例えば，エントリ数を 1K から 16K に増加することによって，PPC ミス率は平均 15.1 ポイント改善する．このように，従来の PPC の容量は十分とは言えず，大容量化によって TCAM アクセスを大いに減らすことができる．

なお，128K エントリ以上の PPC ではどのトレースでもほぼミス率の改善は見られない．これは，PPC が最大でも 128K エントリ程度あれば十分であることを示しており，これ以上のミス削減には初期参照ミスの削減が必要であることを示している．

4. キャッシュのスループット改善手法

4.1 階層化

マイクロプロセッサでは，キャッシュを階層化するアーキテクチャが一般的に採用されている．すなわち，小容量かつ高速なキャッシュを上位レベルに配置し，大容量かつ低速なキャッシュを下位レベルに配置する．

マイクロプロセッサにおいては，キャッシュの階層化の効果を平均メモリアクセスレイテンシの観点から論じることが多く，メモリスループットの観点から論じることは少ないため，注意が必要である．階層化により，高スループットな上位レベルのキャッシュからのデータ供給が増える一方，低スループットな下位レベルのキャッシュあるいはメモリからのデータ供給は減るため，メモリシステム全体のスループットは改善する．

PPC の階層化では，上記の効果を利用してテーブル検索処理全体のスループットを改善する．なお，具体的な効果については，3 階層化によって 1Tbps のパケット処理性能を達成できることが既に明らかとなっている．詳しくは我々の過去の研究報告を参照されたい [13]．

4.2 パイプライン化

パイプライン化は，キャッシュアクセスを複数のステージへ分割することによりサイクル時間を削減する手法である．キャッシュへのアクセスは毎サイクル開始できるため，キャッシュをパイプライン化することによりスループットが向上する．しかし，キャッシュを構成する SRAM 内部に複数のパイプラインレジスタを挿入する必要があるため，パイプライン化によりキャッシュの回路規模や遅延はわずかに増加する欠点がある．

表 1: 探索したキャッシュエントリ数の範囲

	L1	L2	L3
L1 PPC	1/8 K - 128K	N/A	N/A
L2 PPC	1/8 K - 8K	8K - 128K	N/A
L3 PPC	1/8 K - 1/2 K	1K - 16K	8K - 128K

4.3 マルチポート化

マルチポート化はキャッシュへのリードポートを複数設けることによりキャッシュアクセスを並列化する手法である．このとき，スループットは基本的にはポート数に比例して増加する．ただし，ポート数を増やすことによりキャッシュの回路規模は増大する（一般には SRAM の回路面積はポート数の 2 乗に比例して増加する）ため，面積制約を考慮した場合，マルチポート化が適用可能なケースは自ずと限定される．特に大容量のキャッシュではマルチポート化による面積の増加が著しいため，この手法は小容量のキャッシュにおいて有用である．

4.4 上記 3 つの組み合わせ

前節までに述べた 3 つの手法（階層化，パイプライン化，マルチポート化）は直交した関係にあるため，これらの手法を任意の組み合わせで PPC に適用することができる．階層化，パイプライン化，マルチポート化を同時に PPC に適用することにより，それぞれの手法を単独で適用した場合よりも高スループット，省電力，小面積な PPC を実現することが期待できる．

5. 実験手法

4 章で述べた手法を PPC に適用し，さまざまな設計制約において最適な PPC の構成を探索した．PPC ミス率は我々が開発した PPC シミュレータを用いて評価した．[*1]

PPC のエントリ格納方式には，実ハードウェアでの実装を見越し，現実的なコストで実装が可能であるダイレクトマップ方式とセットアソシアティブ方式について評価した．セットアソシアティブ方式においては 2 ウェイおよび 4 ウェイについて評価した．また，PPC 階層間のエントリ制御にはライトスルー方式およびインクルーシブキャッシュを採用した．**表 1** にシミュレーションにおいて評価した PPC の各レイヤにおけるエントリ数を示す．PPC のエントリ数は，1/8K~512K の範囲で変化させた．ただし，複数階層の構成においては，下位のキャッシュの容量が上位のキャッシュ容量よりも大きくなる構成のみを評価した．

PPC のアクセス時間，サイクル時間，消費電力，回路面積の見積もりには CACTI 6.5[10] を用いた．CACTI 6.5 では，ユーザがパイプライン段数を直接指定して SRAM のシミュレーションを行うことはできず，サイクル時間，面積，消費電力等の各パラメータを設計時にどの程度重視するかという重みのみユーザが指定できる．CACTI 6.5 は，この各パラメータの重みを元に，パイプライン段数を含めた SRAM の最適な回路構成を自動的に計算する．実験にあたっては，これらの重みを 25%および 33.3%刻みで変更することにより多くの組み合わせを探索した．得られた見積もり結果より，アクセス時間がサイクル時間より長い，つまりアクセスに 2 サイクル以上を必要とする構成をパイプライン化されているものとして評価した．

[*1] ソースコードは GitHub にて公開している: `https://github.com/kyontan/cache_simulator`

表 2: 評価に用いた CACTI 6.5 のパラメータ

パラメータ	値
プロセステクノロジ	32nm
ポート	1/2/4(リード) + 1(ライト)
トランジスタモデル	ITRS-HP / ITRS-LSTP

評価に用いた CACTI のパラメータを**表 2** に示す．なお，パラメータのうち斜線で複数の値を表記したものについては，それぞれの組み合わせを全て探索した．

階層化した PPC におけるスループット T_{PPC}[Gbps] の計算式を，2 階層の場合を例に以下に示す．なお，以下の式は 3 階層以上の PPC のスループットを計算する場合にも容易に拡張できる．

$$T_{\mathrm{PPC}} = \min\left(\frac{l \cdot p_{\mathrm{L1}}^{read}}{d_{\mathrm{L1}}}, \frac{l \cdot p_{\mathrm{L2}}^{read}}{d_{\mathrm{L2}} \cdot m_{\mathrm{L1}}}, \frac{l}{d_{\mathrm{TCAM}} \cdot m_{\mathrm{L1}} \cdot m_{\mathrm{L2}}}\right) \quad (1)$$

ここで d_{Ln}，d_{TCAM} はそれぞれ Ln PPC，TCAM のサイクル時間を表す．下位の PPC のクロックには最上位の PPC のクロックを分周したものが入力されることを想定し，各階層の PPC のサイクル時間が L1 PPC のサイクル時間の定数倍となるように切り上げた．なお，d_{TCAM} については論文 [2] を参照した．m_{Ln}, p_{Ln}^{read} はそれぞれ Ln PPC のミス率とリードポート数を表す．ここで，Ln は PPC の n 層目を表す．l は 1 パケットのサイズであり，本報告では最悪ケースを想定し，IP における最小パケット長である 64bytes とした．

一方，2 階層の PPC における平均消費電力 P_{PPC}[mW] は次式で表される．下記の式も 3 階層以上の PPC の平均消費電力の計算式に容易に拡張可能である．

$$\begin{aligned} P_{\mathrm{PPC}} = &\left(E_{\mathrm{L1}}^{\mathrm{dynamic}} + E_{\mathrm{L2}}^{\mathrm{dynamic}} \cdot m_{\mathrm{L1}}\right) \cdot n \\ &+ \left(a \cdot E_{\mathrm{TCAM}}^{\mathrm{dynamic}} \cdot m_{\mathrm{L1}} \cdot m_{\mathrm{L2}}\right) \cdot n \\ &+ P_{\mathrm{L1}}^{\mathrm{static}} + P_{\mathrm{L2}}^{\mathrm{static}} + P_{\mathrm{TCAM}}^{\mathrm{static}} \end{aligned} \quad (2)$$

$E_{\mathrm{Ln}}^{\mathrm{dynamic}}$，$E_{\mathrm{TCAM}}^{\mathrm{dynamic}}$ はそれぞれ Ln PPC と TCAM の動的消費エネルギーを表している．また，$P_{\mathrm{Ln}}^{\mathrm{static}}$，$P_{\mathrm{TCAM}}^{\mathrm{static}}$ はそれぞれ Ln PPC と TCAM の静的消費電力を表している．ここで，$E_{\mathrm{TCAM}}^{\mathrm{dynamic}}$ および $P_{\mathrm{TCAM}}^{\mathrm{static}}$ に関しては，ルータで一般的に用いられている 20Mbit TCAM を想定し，電力は容量に比例するものと仮定して論文 [2] の値を定数倍することで見積もった．a は 1 パケットのテーブル検索処理に要

する TCAM アクセス数である．本報告では図 1 に示した一般的なインターネットルータを想定し，$a = 4$ とした．n は 1 秒あたりの平均パケット数を表している．

評価には，NLANR AMP [12] および WIDE MAWI Working Group Traffic Archive [14] より取得した 10 種の実ネットワークにおけるトレースを使用した．[14] は WIDE MAWI Working Group Traffic Archive から取得したものであり，2017 年に WIDE と上流 ISP を接続する 1Gbps のリンクから取得された 15 分間のトレースである．消費電力については，1Gbps のリンクである WIDE のネットワークトレースにおける 1 秒あたりの平均パケット数を元に，設計制約で求めるスループットに対して定数倍することにより，スループットの増大に比例してパケット数が増加すると仮定して計算を行った．

面積評価では，インターネットルータのチップ内において大部分を占めるとされている [7] パケットバッファと PPC のみを評価対象とした．インターネットルータのパケットバッファの容量は，商用の 100Gbps ルータ [3] におけるオンチップパケットバッファを参考とし 16MB とした．以降では，**PPC の面積**と表記した場合にはパケットバッファの面積は含まないものとする．

6. 実験結果

以降の説明では，従来の PPC とは 1 階層, 1K エントリ，1 リードポートからなる PPC を指すものとする．従来の PPC の見積もりにおいて CACTI に与えた重みは初期パラメータであり，パイプライン化されていない SRAM の見積もり結果となっている．この構成におけるスループット，消費電力，面積はそれぞれ 294.8Gbps，151.7mW および 29.80mm^2 であった．なお，面積制約を考える際は，従来の PPC のパケットバッファを含む面積である 29.80mm^2 を 100%とする．また，簡単のため，マルチポート化とパイプライン化のみを適用した構成を**組み合わせ (M+P)**，これに加え，更に n 階層の階層化を適用した構成を**組み合わせ (Hn+M+P)** と表す．

6.1 スループットおよび消費電力，面積の分布

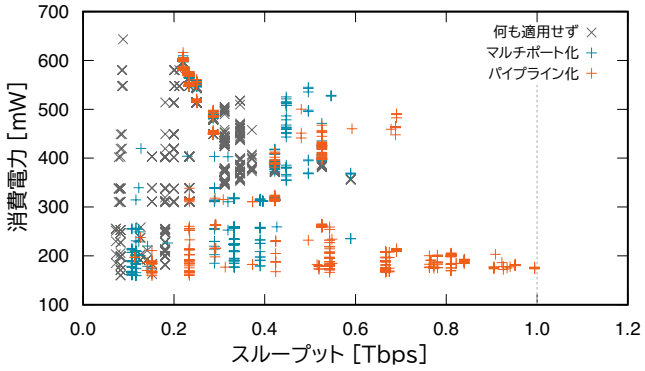
まず，各手法を個別に，また組み合わせて適用した場合のスループットと消費電力，PPC の面積について，様々なメモリ構成で見積もった．結果はランダムサンプリングにより各手法ごとに 500 件へ絞った上で，散布図にプロットしている．また，これに加え，特筆して説明すべき構成についてもプロットした．

1 階層の PPC においてパイプライン化とマルチポート化をそれぞれ適用した場合の，各メモリ構成におけるスループットと消費電力，PPC の面積の分布を**図 4** に示す．どの手法も適用していない場合に比べ，パイプライン化またはマルチポート化を適用することでスループットの向上が見られた．特に，パイプライン化によるスループット向上は顕著であり，パイプライン化のみを適用した場合でも 1Tbps に近いスループットを達成できることが明らかになった．また，消費電力の観点から見ても，マルチポート化ではスループットの向上に伴い消費電力が増加する傾向にある一方で，パイプライン化では消費電力の増加が見られなかった．PPC の面積については，800Gbps より高いスループットを達成するためには，スループットの向上に伴い大きい面積の PPC が必要となることが分かった．特に，パイプライン化を適用した最もスループットの高い構成では，最小でも面積が 52.50mm^2 となり，従来の PPC と比べ面積が 76.2%増大している．

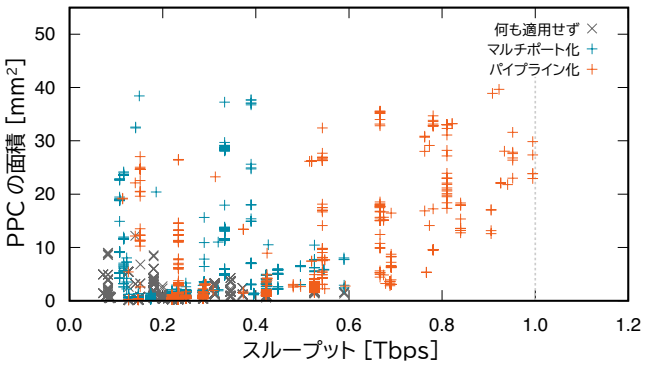
次に，階層化のみを適用した場合の，各メモリ構成におけるスループットと消費電力，PPC の面積の分布を**図 5** に示す．階層化については過去の研究報告 [13] にて述べた通りであるが，2 階層では 1Tbps を達成する構成はなく，3 階層では 1Tbps のパケット処理スループットを達成する構成がいくつか存在する．階層化を適用した構成では，定義より最小の ETLT を持つメモリにスループットが律速されるため，同一のスループットで異なる面積や消費電力を示す構成が多く見られた．階層化した構成で最もスループットが高く，面積が最小となった構成ではスループット，消費電力および面積はそれぞれ，1.05Tbps，386.9mW，および 58.62mm^2 となった．これは，従来の PPC に比べてスループットを 3.55 倍向上し，消費電力を 14.5%削減する一方で，96.7%の面積増大を招いている．

最後に，3 手法を組み合わせた場合の結果を**図 6** に示す．3 手法を組み合わせた場合，2 階層の PPC においても 1Tbps を達成する構成が存在することが明らかになった．1Tbps を超える構成のうち，面積の最小化を設計制約とした場合には，組み合わせ (H2+M+P) と組み合わせ (H3+M+P) の構成では面積がそれぞれ 49.78mm^2 および 49.73mm^2 となり，組み合わせ (H3+M+P) の方が僅かに面積が少なくなる結果となった．これらの構成は，どちらも 4K エントリおよび 128K エントリの PPC を用いているが，組み合わせ (H3+M+P) の構成では前段により小さな 1/8K エントリの PPC を用いている．これにより，L2 PPC へのアクセス率を 42.9%まで削減し L2 PPC のパイプライン化を回避することで面積が削減されたとみられる．

同様に，1Tbps を超える構成のうち電力最小化を設計制約とした場合でも，組み合わせ (H2+M+P) と組み合わせ (H3+M+P) の構成では消費電力がそれぞれ 143.1mW および 137.4mW となり，組み合わせ (H3+M+P) の構成の方が消費電力が若干少ない結果となった．これらは PPC の構成が大きく異なるが，面積の最小化の場合と同様に小容量のキャッシュを前段に用いることで後段の大容量 PPC へのアクセスを削減することにより消費電力を削減できていると考えられる．

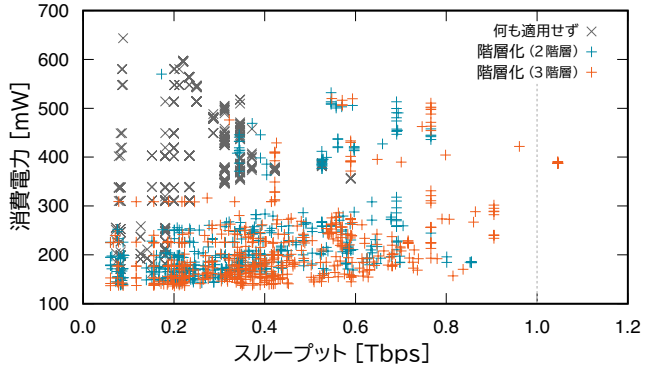


(a) スループットと消費電力

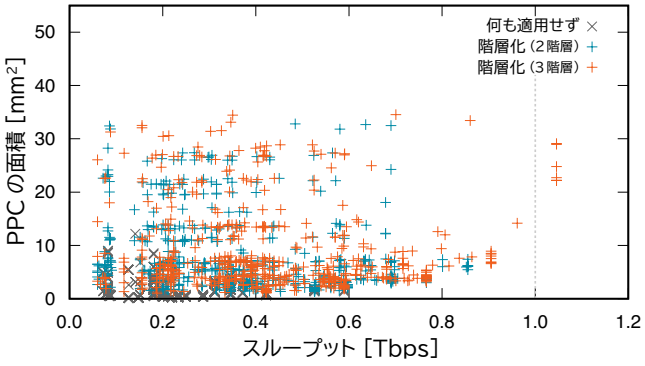


(b) スループットと PPC の面積

図 4: パイプライン化とマルチポート化をそれぞれ適用した場合のスループットと PPC の面積，消費電力の比較

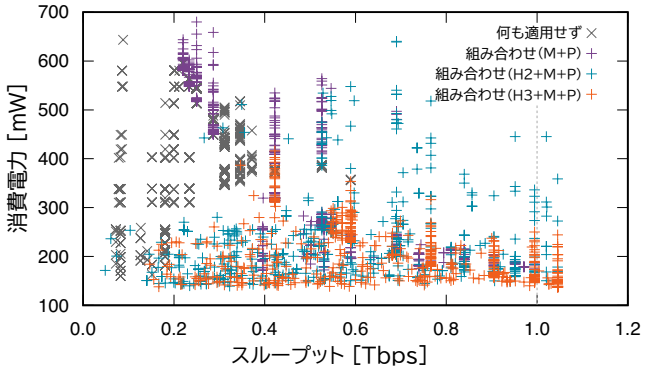


(a) スループットと消費電力

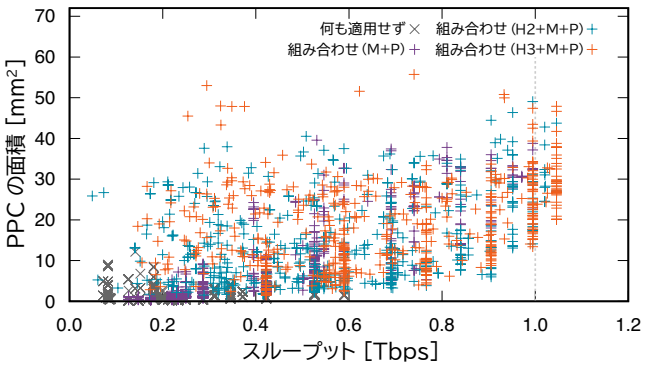


(b) スループットと PPC の面積

図 5: パイプライン化とマルチポート化をそれぞれ適用した場合のスループットと PPC の面積，消費電力の比較



(a) スループットと消費電力



(b) スループットと PPC の面積

図 6: 各手法を組み合わせて適用した場合のスループットと PPC の面積，消費電力の比較

6.2 各手法の比較

各手法の特徴を詳しく分析するため，どの手法でも達成可能であるスループット 500Gbps 以上，面積増加 20%以下，電力最小化を設計制約として各手法ごとに最適な構成を特定した．まず，**図 7** にメモリアクセスの内訳を示す．TCAM へのアクセス率は，マルチポート化した構成では比較的高い値を示したが，その他の構成では大きな差は見られなかった．一方で，2 階層や 3 階層の構成では手法を組み合わせた場合とそうでない場合とで，各階層へのアクセス率が大きく異なる結果となった．多層化のみを行った構成では L1 PPC を小容量化することで L1 PPC の動的消費電力を削減している．一方で，各手法を組み合わせた場合には，組み合わせ (H2+M+P) では L2 PPC へのアクセス率が増大し，組み合わせ (H3+M+P) では減少する結果となった．これは次に述べる消費電力の最小化が大きく関係している．

図 8 に消費電力の内訳を示す．消費電力の内訳を見ると，マルチポート化を除いては各手法の適用による消費電

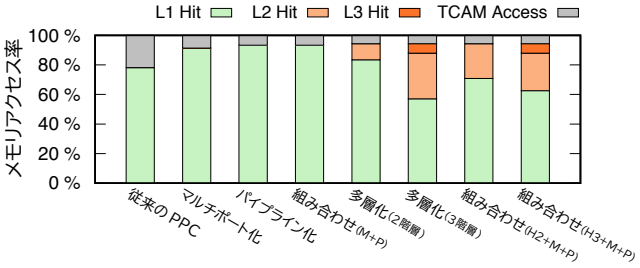


図 7: スループット 500Gbps 以上，面積増加 20%以下で消費電力の最小化を設計制約としたときの各手法におけるメモリアクセスの内訳

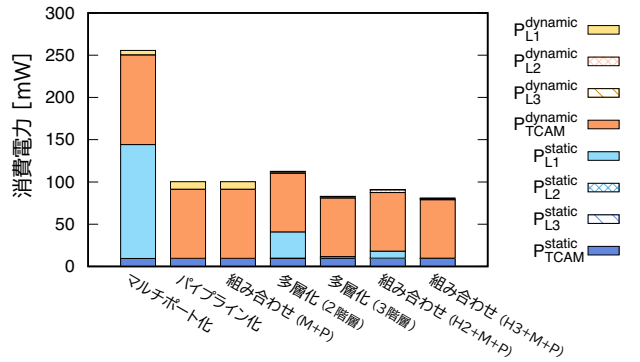


図 8: スループット 500Gbps 以上，面積増加 20%以下で消費電力の最小化を設計制約としたときの各手法における消費電力の内訳

力の削減効果に大きな差はないことが分かる．各手法を組み合わせた構成では 2 階層，3 階層ともにパイプライン化と階層化のみを適用した構成が最適となった．これはパイプライン化により 2 階層では L1 および L2 PPC を，3 階層では L1 PPC をパイプライン化することにより，より柔軟にエントリ数とサイクル時間の構成を変化させることが可能になったためであると推測する．また，マルチポート化は顕著に静的消費電力が増加しているが，これはマルチポート化のみが影響しているわけではないことに注意する必要がある．この構成は高速，高消費電力な ITRS-HP トランジスタモデルを用いており，低消費電力の ITRS-LSTP トランジスタモデルでは設計制約を満たすことができなかった．そのため，静的消費電力が極めて大きくなっている．ITRS-HP トランジスタモデルは，2 階層化を行った構成における L1 PPC でも用いており，こちらも同様に静的消費電力が他の構成と比べて大きくなっている．

6.3 面積制約を変えることによる消費電力の比較

図 9 に，組み合わせ (H3+M+P) でスループット 1Tbps 以上を目的とし，面積制約を変えたときの消費電力を示した．ここで，図中の $P^{\mathrm{dynamic}}_{\mathrm{Ln}}, P^{\mathrm{dynamic}}_{\mathrm{TCAM}}$ はそれぞれ Ln PPC と TCAM の動的消費電力であり，動的消費エネルギーに想定するパケット数を乗じることにより得たもので

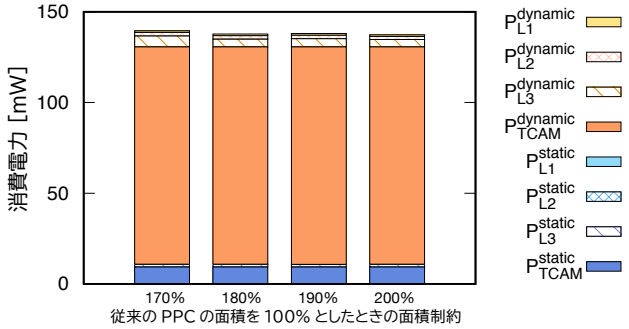


図 9: 組み合わせ (H3+M+P) において，スループットが 1Tbps を超える構成で面積制約を変更したときの消費電力の内訳

ある．図から明らかなように，面積制約を変えても消費電力への影響は見られなかった．

1Tbps のスループットを達成するためには 128K エントリの LLC が必須であり，2 章で説明したようにこれ以上エントリ数を増やしても TCAM へのアクセス率削減は見込めない．そのため，面積制約を緩和し，L3 PPC の容量を除く構成の自由度が高くなった状況においても，消費電力は大きく削減できなかった．つまり，TCAM の動的消費電力が PPC の消費電力を支配している状況下で，面積制約を変えることによる消費電力の影響は軽微であることが分かった．

6.4 要求スループットを変えることによる消費電力の比較

図 10 に，要求するスループットをそれぞれ 600Gbps, 800Gbps，1Tbps とし，面積制約としてそれぞれ 120%, 120%, 170%を与え，消費電力を最小化したときの消費電力の推移を示す．これらの面積制約は，各スループット制約を満たしつつ面積が最小となる構成の面積の近傍で消費電力がより少なくなる構成を探索するために設定したものである．各スループットにおいて，この面積制約より 10%以上面積が少ない構成は存在しない．

図に示したように，面積制約が 120%である 2 つの構成では，スループットに比例して TCAM の動的消費電力が増大していることが分かる．一方で，スループットの設計制約が 1Tbps，面積制約が 170%の構成では TCAM の動的消費電力の増加度合いが少ないことが分かる．これはラストレベルキャッシュ (LLC) の容量差が影響している．LLC は，面積制約が 120%の場合には最大 32K エントリ，170%の場合には最大 128K の構成を取ることができ，消費電力は TCAM の動的消費電力が支配的であるため，消費電力を最小化しようとした場合には必然的に LLC の容量を最大化し，TCAM へのアクセス率を削減する必要がある．よって，消費電力削減の観点からは要求するスループットが低い場合にも大容量の LLC を組み入れることが望ましいことが分かる．

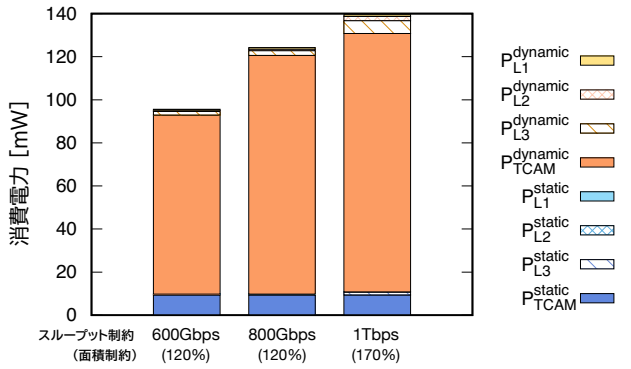


図 10: 組み合わせ (H3+M+P) において，スループットと面積を設計制約として与えたときの消費電力の内訳

6.5 最適な構成の検討

本研究では，インターネットルータのパケット処理スループットで 1Tbps を達成すると同時に，面積および消費電力の削減を主な目的としている．これらを鑑み，1Tbps のスループットを面積増加 70%以下で達成し，最も消費電力が少ない構成を最適な構成として選択し，ETLT とメモリアクセスの内訳について分析を行った．

これらの設計制約を満たすことができるのは組み合わせ (H2+M+P) と組み合わせ (H3+M+P) による構成のみであり，組み合わせ (H3+M+P) による構成の方が消費電力が少ないため，これを元に分析を行った．**表 3** に構成と性能の要約を示す．ここで，従来の PPC の消費電力は，その構成達成できる 294.8Gbps のスループットを仮定したものであることに注意されたい．階層化のみを適用した場合，1Tbps のスループットを達成する構成の中で消費電力を最小化したときに達成できる消費電力は 386.9mW であった．これに比べ，組み合わせ (H3+M+P) は 63.9%少ない消費電力で 1Tbps を達成することができており，1Gbps あたりの消費電力は 74.1%改善した．

図 11 に ETLT の分析結果を示す．依然として MRA, ODU, WIDE の 3 トレースで TCAM の ETLT が 1Tbps に満たない結果となったが，この結果はこれらのトレースが時間的局所性の低いネットワークで取得されたトレースであることを顕著に反映しており，マルチポート化やパイプライン化の導入により改善が見込めるものではない．特に，UPCB，MRA，ODU および WIDE は 図 3 からも分かるように PPC のエントリ数を 128K エントリまで増加させても TCAM へのアクセス率が 1Tbps の要求水準に達しない．つまり，これらのトレースに対して 1Tbps を実現するためには初期参照ミスの削減が必須であることを示している．

図 12 にメモリアクセス率の内訳を従来の PPC と比較して示した．ネットワークトレースにより PPC の効果は大きく異なるが，平均して TCAM アクセス率は 77.7%削減された．

表 3: 従来の PPC とスループット 1Tbps 以上，70%以下の面積増加で消費電力最小を設計制約として与えたときの構成と性能の要約

		BASE	組み合わせ (H3+M+P)
エントリ数		1K	1/4K / 4K / 128K
ウェイ数		4	2 / 4 / 4
リードポート数		1	2 / 2 / 1
パイプライン段数		1	2 / 1 / 2
サイクル比		N/A	1 : 5 : 21
トランジスタモデル		ITRS-HP	ITRS-LSTP (全階層)
スループット		294.8Gbps	1045.7Gbps
消費電力		151.7mW	139.5mW
合計面積		29.80mm^2	50.48mm^2
面積	L1 PPC	0.22mm^2	0.18mm^2
	L2 PPC	N/A	1.19mm^2
	L3 PPC	N/A	19.53mm^2
パケットバッファ		29.58mm^2	29.58mm^2

/ で区切られた値は左から順に L1, L2, L3 PPC の値を表している

7. 関連研究

PPC におけるキャッシュミスは初期参照ミス，容量性ミス，そして衝突性ミスに分類され，それぞれに対して様々な改善手法が提案されてきた．

容量性ミスに着目した研究として，Cheng らはキャッシュタグを圧縮することによるエントリあたりのサイズの削減手法を検討している [1]．ルータ内の各テーブルで用いられるフロー情報は 104bit と大きいことから，32bit のハッシュ値をタグ情報として用いる Digest Cache を提案している．Digest Cache はブルームフィルタをルータで用いるために改変した手法であり，許容される割合の誤ったテーブル検索結果と引き換えに高い容量あたりのエントリ数を実現している．

衝突性ミスに焦点を当てた研究として，Kim らは過去 2 回のパケットのタイムスタンプの合算値が小さいエントリを追い出す手法である L2A scheme を提案している [9]．Kim らは一般的に使用されるエントリ置換アルゴリズムである LRU は時間的局所性のみを考慮しており，インターネットルータで考慮すべきネットワークの特性を反映しておらず高い効果を発揮できないと主張している．L2A scheme は同論文内で提案されているもう 1 つの提案手法である Weighted Priority LRU scheme に比べ長期間に渡り複数回参照されたエントリを保持し，キャッシュの平均置換回数が優れているとした．しかしながら，タイムスタンプを保持するためのメモリやハードウェアのコストには触れられておらず，メモリ容量の少ない PPC ではメモリコストの増大が致命的な問題になると考えられる．

我々も，過去に PPC のミス削減手法として，フローの構成パケット数に着目した新たなエントリ置換方式の検討 [16]，初期参照ミスの削減を目的としたキャッシュエン

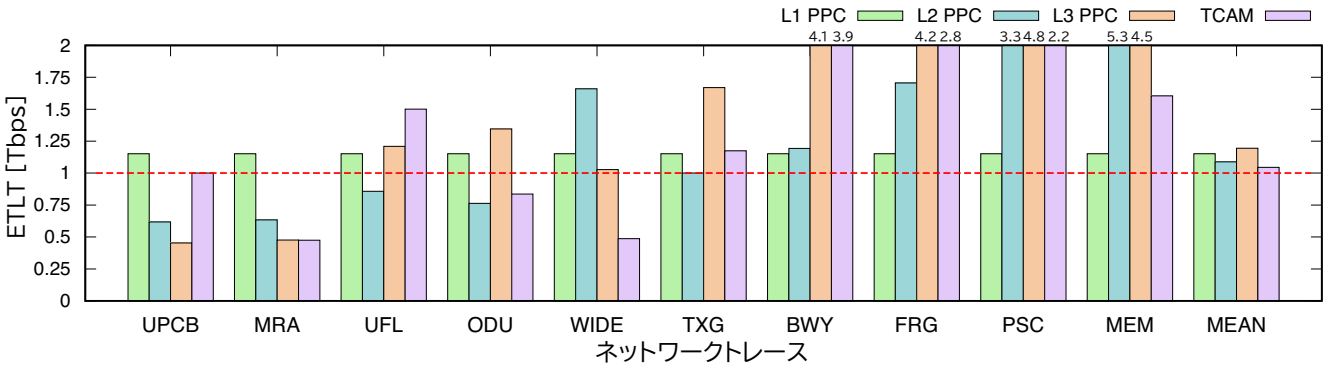


図 11: スループット 1Tbps 以上，面積増加が 70%以下で消費電力が最小の構成における ETLT

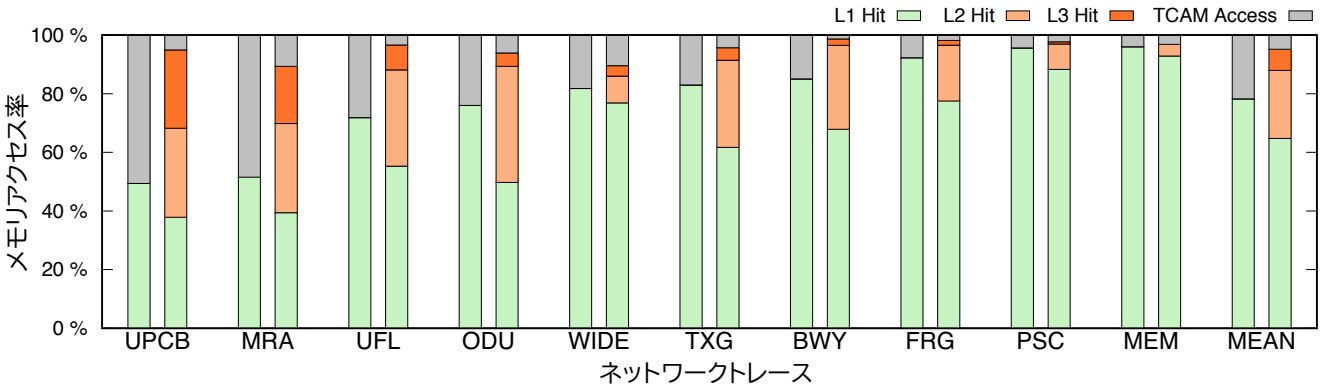


図 12: スループット 1Tbps 以上，面積増加が 70%以下で消費電力が最小の構成におけるメモリアクセス率の内訳

トリの挿入方式の検討を行った [15]．論文 [16] では，パケット数の多いフローのエントリを長期間保持し，少数のパケットからなるフローのエントリを迅速に追い出すエントリ置換アルゴリズムとして ELC (Elevator Cache) を提案した．これにより，PPC のミス率を LRU と比べて平均 11.1%，削減している．また論文 [15] では，初期参照ミスの削減のため，トラフィックにサーバ・クライアント型の通信が多いことに着目した投機的なキャッシュの挿入方式を提案している．実ネットワークにおけるトラフィックでは，およそ 70%が対称なフロー，つまり送信元と送信先の IP アドレスとポートを入れ替えたフローから構成されている．この特性を利用し，フローの最初のパケットを処理するとき，その後に対称なフローのパケットが来ることを期待して対称なフローのエントリを投機的にキャッシュする RPC (Response Prediction Cache) を提案し，キャッシュミスを従来の PPC に比べ平均 15.3%削減している．

8. 結論

本研究では，インターネットルータのスループット向上および省電力化を実現するパケット処理キャッシュ (PPC) に対して，階層化，パイプライン化，マルチポート化，あるいはその組み合わせを適用し，これらの手法が PPC に与える影響について，包括的に評価した．評価の結果，3 つの手法を組み合わせ，従来の PPC から 70%の面積増加を許容することにより，従来の PPC より 8.0%少ない消費電力でスループットを 255%向上し，1Tbps を達成する構成を特定した．

謝辞 本研究は，JSPS 科研費（JP18K18022）による助成を受けたものである．

参考文献

[1] Chang, F., Feng, W., Feng, W. and Li, K.: *Network Processor Design, Volume 3: Issues and Practices (The Morgan Kaufmann Series in Computer Architecture and Design)*, chapter Efficient Packet Classification with Digest Caches, pp. 33–54, Wiley (2005).

[2] Cheng, Y., Chen, J., Wu, T. and Chang, Y.: Low Leakage Mask Vertical Control TCAM for Network Router, *Proc. of the 2016 IEEE Asia Pacific Conference on Circuits and Systems*, APCCAS'16, pp. 469–472 (online), DOI: 10.1109/APCCAS.2016.7804005 (2016).

[3] Cisco: Cisco Nexus 9500 R-Series Line Cards and Fabric Modules White Paper, Cisco (online), available from ⟨https://www.cisco.com/c/en/us/products/collateral/switches/nexus-9000-series-switches/white-paper-c11-738392.html⟩ (accessed 2019-6-21).

[4] Cisco: Cisco Visual Networking Index: Forecast and

Trends, 20172022 White Paper, Cisco (online), available from ⟨https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-741490.html⟩ (accessed 2019-6-11).
[5] Eatherton, W., Varghese, G. and Dittia, Z.: Tree Bitmap: Hardware/Software IP Lookups with Incremental Updates, *SIGCOMM Comput. Commun. Rev.*, Vol. 34, No. 2, pp. 97–122 (online), DOI: 10.1145/997150.997160 (2004).
[6] Ethernet Alliance: The 2019 Ethernet Roadmap, Ethernet Alliance (online), available from ⟨https://ethernetalliance.org/the-2019-ethernet-roadmap/⟩ (accessed 2019-6-12).
[7] Hassan, H., Said, M. and Kim, H.: *Proc. of the 59th IEEE International Midwest Symposium on Circuits and Systems*.
[8] Hewlett Packard Enterprise: Energy Efficient Networking - Business White Paper, Hewlett Packard Enterprise (online), available from ⟨http://h17007.www1.hpe.com/docs/mark/4AA3-3866ENW.pdf⟩ (accessed 2019-6-21).
[9] Kim, N., Jean, S., Kim, J. and Yoon, H.: Cache replacement schemes for data-driven label switching networks, *2001 IEEE Workshop on High Performance Switching and Routing (IEEE Cat. No.01TH8552)*, pp. 223–227 (online), DOI: 10.1109/HPSR.2001.923636 (2001).
[10] Muralimanohar, N., Balasubramonian, R. and Jouppi, N.: Optimizing NUCA Organizations and Wiring Alternatives for Large Caches with CACTI 6.0, *Proc. of the 40th Annual IEEE/ACM International Symposium on Microarchitecture*, MICRO-40, pp. 3–14 (online), DOI: 10.1109/MICRO.2007.30 (2007).
[11] Nawa, M., Okuda, K., Ata, S., Kuroda, Y., Yano, Y., Iwamoto, H., Inoue, K. and Oka, I.: Energy-efficient high-speed search engine using a multi-dimensional TCAM architecture with parallel pipelined subdivided structure, *13th IEEE Annual Consumer Communications & Networking Conference, CCNC 2016, Las Vegas, NV, USA, January 9-12, 2016*, pp. 309–314 (online), DOI: 10.1109/CCNC.2016.7444794 (2016).
[12] Réseaux IP Européens Network Coordination Centre: NLANR AMP Data, , available from ⟨https://labs.ripe.net/datarepository/data-sets/nlanr-amp-data⟩ (accessed 2019-6-21).
[13] Tanaka, K., Yamaki, H., Miwa, S. and Honda, H.: Multi-Level Packet Processing Caches, *2019 IEEE Symposium in Low-Power and High-Speed Chips (COOL CHIPS)*, pp. 1–3 (online), DOI: 10.1109/CoolChips.2019.8721336 (2019).
[14] WIDE MAWI Working Group: MAWI Working Group Traffic Archive, , available from ⟨http://mawi.wide.ad.jp/mawi/⟩ (accessed 2019-6-21).
[15] Yamaki, H., Nishi, H., Miwa, S. and Honda, H.: Data Prediction for Response Flows in Packet Processing Cache, *Proc. of the 55th Annual Design Automation Conference*, DAC '18, pp. 110:1–110:6 (online), DOI: 10.1145/3195970.3196021 (2018).
[16] Yamaki, H. and Nishi, H.: Line Replacement Algorithm for L1-scale Packet Processing Cache, *Adjunct Proceedings of the 13th International Conference on Mobile and Ubiquitous Systems: Computing Networking and Services*, MOBIQUITOUS 2016, New York, NY, USA, ACM, pp. 12–17 (online), DOI: 10.1145/3004010.3006379 (2016).